## Remarks/Arguments:

Claims 15-38, previously presented, with claims 15-18 amended, hereby, are pending.

The amendments to claims 15-18, effected hereby, are discussed in the context of addressing the rejections of record, below.

The Office Action contains three rejections, each of which is applied against all pending claims, i.e., claims 15-38: (1) under §112, first paragraph, for allegedly containing new matter; (2) under §112, second paragraph, for allegedly being indefinite; and (3) under §112, first paragraph, for allegedly lacking enablement.

In response to the new-matter rejection, under 35 USC 112, first paragraph, the phrase "of which the melting initiation point and the mobility transition end point are known" – allegedly constituting new matter – is deleted from claims 15, 16, 17 and 18 by the instant amendment. Accordingly, the rejection appears to be in order for withdrawal.

In response to the rejection under 35 USC 112, 2nd ¶, each occurrence of the allegedly indefinite terminology – "Eq." – is changed to read "equation." Accordingly, the rejection appears to be in order for withdrawal.

Reconsideration is requested with respect to the rejection of claims 15-38 under 35 USC 112, first paragraph, for allegedly lacking enablement.

The rejection is based on the allegation that the presently claimed method would not work for its intended purpose (i.e., enablement of *use* rejection). Essentially, it is argued that the presently claimed method would result in 98% false positive results (Office Action, page 5, first incomplete

paragraph). In coming to this conclusion, the statement of rejection alleges (Office Action, page 4, last incomplete paragraph) (emphasis added):

> Random PCR is reasonably expected to amplify both conserved *and* variable sequences in a genome. It is reasonably expected that conserved sequence analysis will always result in an indication of 100% similarity between organisms being compared and thus contribute to a false positive result. The variable sequence amplification products may result in indicating sequence differences however these are very low percentage of sequences in reasonably related organisms.

Put another way, it is alleged that

a) a person skilled in the art would have expected use of PCR, in the manner presently claimed, to effect amplification of "both conserved and variable sequences," in the amplified fragments of target-organism-genome DNA,

b) the presently claimed method involves "conserved sequence analysis," and

c) "Conserved sequence analysis will always result in an indication of 100% similarity between organisms being compared and thus contribute to a false positive result."

Applicant submits that these allegations result from an apparent misunderstanding; i.e., that conserved sequences are completely identical in similar organisms without considering point mutations in the conserved sequences.

The statement of rejection maintains that, not only *variable region* sequences but, also, *conserved region* sequences contain variations in nucleic acid sequence. It is true that the degree of variation in the variable region is much bigger than that in the conserved region; however, not all sequences in the conserved region of one organism – e.g., human – are identical to those in a similar

organism – e.g., chimpanzee, consistently. This lack of consistent identity is, of course, the result of random *mutation*. Even in the same organism there are variations in conserved sequences due to mutation. In accordance with the presently claimed invention, the skilled artisan is, now, enabled to identify the species of a *target* organism by "determining the similarity at the genome level between [the] target organism and a reference organism" (claim 15).

For the Examiner's convenience, Applicant encloses herewith a copy of *Gene 261* (2000) 243-250, co-authored by one of the inventors of the subject application, Dr. Koichi Nishigaki. In the *Gene* article, the authors explain that variations in conserved sequences exist due to mutation; and, the variations occur even among strains classified in the same species (*E. coli* O157:H7); and, further, the strains can be distinguished by the method of the present invention – identified as "spiddos method" in the *Gene* article.

Specifically, the *Gene* article reports that 19 strains were examined. Genome profiling of each strain is shown in Fig. 2A the *Gene* article (page 246) and data of PaSS of every two stains are summarized in Fig. 2B (page 247). In addition, sequences of $\alpha$, $\beta$, and $\gamma$ indicated in Fig. 2A were determined and shown in Fig. 4B. As seen from Fig. 4B of the *Gene* article (page 248), sequences in these three regions are conserved very well, but they contain variation in sequence due to mutation. This means that, even among organisms of the same species, conserved regions contain some mutations – in other words, conserved regions contain sequence *variation*. By means of this variation in sequence, the *Gene* article demonstrates that the method disclosed and claimed in the subject application – exemplified by the *spiddos method* disclosed in the *Gene* article – enables one

15

skilled in the art to use the claimed invention – to distinguish between strains even though the strains are classified in the same species – and, so, evidences that the requirements for enablement under §112, ¶1, are satisfied. Accordingly, withdrawal of the rejection appears to be in order.

Applicant is not unmindful of the point raised in the statement of rejection in reliance on US6228586B. US6228586B (Fig. 2) discloses a comparison between the nucleic acid sequence of *intercellular adhesion molecule*-1 ( ICAM-1) in a human and the nucleic acid sequence of ICAM-1 in a chimpanzee. Since the sequences are those of (ICAM-1, they should represent what is the statement of rejection alleges to be a "conserved sequence." However, as shown in US6228586B Fig. 2 (copy attached), comparing the first 480 nucleic acids of human ICAM-1 and chimpanzee ICAM-1, there are 9 pair of nucleic acids in the human that are different from those in the chimpanzee.

US6228586B actually evidences §112 enablement of the presently claimed invention. US6228586B shows that, as between similar species, there are variations in the conserved sequences of their respective genomes. Accordingly, one skilled in the art is enabled to identify a *target* organism by comparison with a *reference* organism in accordance with the method presently claimed.
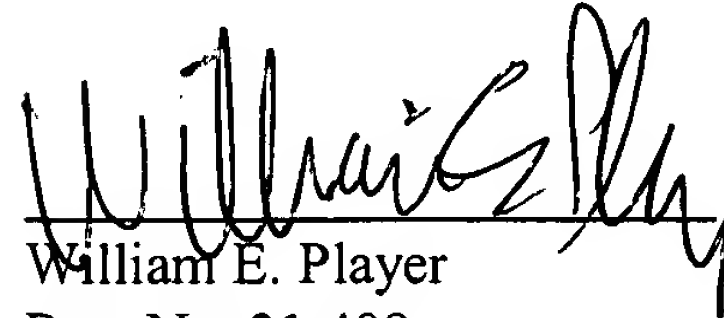
For the Examiner's convenience, Applicant prepared, and attaches hereto, a glossary ("Definitions") of art-recognized terms and phrases terms.

Favorable action is requested.

Respectfully submitted,

JACOBSON HOLMAN PLLC

By     William E. Player
Reg. No. 31,409

400 Seventh Street, NW
The Jenifer Building
Washington, D.C.  20004
Tel. (202) 638-6666
Fax (202) 393-5350
Date: September 20, 2004
WEP/bap
R:\Home\RTHOMAS\2003\OCTOBER\P66602US0 amd.wpd

# Species-identification dots: a potent tool for developing genome microbiology

Mohammed Naimuddin[a], Takayuki Kurazono[b], Yinghua Zhang[c], Takehiro Watanabe[a], Masanori Yamaguchi[b], Koichi Nishigaki[a,*]

[a]*Department of Functional Materials Science, Saitama University, 255 Shimo-Okubo, Urawa, Saitama, 338-8570 Japan*
[b]*Hygiene Institute of Saitama, Kami-Okubo Urawa, Saitama, 338-0824 Japan*
[c]*Department of Scientific Instruments, Taitec Co., Nishikata, Koshigaya-shi, Saitama, 343-0822 Japan*

### Abstract

Identification of species has long been done by phenotype-based methodologies. Recently, genotype-based species identification has been shown to be possible by way of *Genome profiling*, which is based on a temperature gradient gel electrophoresis (TGGE) analysis of random PCR products. However, the results, though sufficient in information, provided by *genome profiling* were complicated and difficult to deal with objectively. To cope with this, a technology of utilizing *species identification dots* (*spiddos*), which corresponds to structural transition points of DNAs, was introduced. *Pattern similarity score* (*PaSS*), derived from *spiddos*, was shown to be usable for quantitatively measuring the closeness between genomes. This was demonstrated with the experiments applied to the genomes of *Escherichia coli* O157:H7 (19 strains). The same genomes were also examined by sequencing and RFLP methods in order to compare the effectiveness of these three methods. As a result, the *spiddos* method was shown to give reasonable results and to be the most advantageous for measuring the closeness between species in general. This means that *spiddos* is pushing the heavy gate open for genome microbiology. © 2000 Elsevier Science B.V. All rights reserved.

*Keywords*: Genome profiling; Pattern similarity score; Genome distance; Genome sequence space; Structure stability

## 1. Introduction

The genotype-based identification of species is more determinative and universal than that based on phenotype. This is particularly true of microbes that generally have few features and are difficult to observe, often impossible to cultivate and exhibit aberrant behaviors under different conditions. However, technological issues have often made such identification unrealistic (Amann et al., 1995). Lately, whole genome sequencing projects have rendered this notion more plausible, although mass sequencing cannot be readily applied to all species. We recently demonstrated species identification using *Genome profiling* without mass sequencing (Nishigaki et al., 2000a). Although *Genome profiling* is highly informative, it is less manageable due to the complexity of the generated data (Nishigaki

et al., 2000a; Nishigaki et al., 1991; Hamano et al., 1996). We examined this issue in more detail in order to allow 'a shift' from phenotype- to genotype-based microbiology, in other words, to establish a new discipline, genome microbiology, which would include for example, genome epidemiology, genome ecology and genome environmental chemistry. The methodology developed here by introduction of species identification dots enabled us to measure the similarity of genomes.

## 2. Materials and methods

### 2.1. Genomic DNAs and primers

*Escherichia coli* O157:H7 genomic DNAs were obtained from Hygiene Institute of Saitama (Saitama, Japan) and those of *Saccharomyces cerevisiae* were prepared from commercial sources using alkaline lysis (Wang et al., 1993).

The primers used were pfM12 (dAGAACGCGCCTG) for random PCR, α-1 (dAGAACGCGCCTGCCTGCG-

---

CAGTAT) / α-2 (dAGAACGCGCCTGAAGTTTATCA-AT) for amplifying α fragment, β-1 (dAGAACG-CGCCTGCCACCACTCGAT) / β-2 (dAGAACGCGCC-TGATCGGAAATAAA) for β fragment and γ-1 (dAGA-ACGCGCCTGTTGCTGGAAGAG)/ γ-2 (dAGAACGC-GCCTGTTCTTCTGATGT) for γ fragment.

## 2.2. Genome profiling

Genome profiling is essentially TGGE analysis of random PCR products. This technology is applicable to all kinds of species (Nishigaki et al., 2000a). A mixture for random PCR contained 10 ng of template DNA, 50 pmoles of primer DNA, 250 μM of each dNTP (N = A, G, C, T), 50 mM Tris–HCl (pH 8.8), 15 mM $(NH_4)_2 SO_4$, 10 mM $MgCl_2$, 0.45% Triton X-100, 200 μg/ml BSA and two units of Taq DNA polymerase (Biotech International). PCR was performed in 30 cycles of 30 s at 94°C, 2 min at 28°C and 2 min at 47°C using a thermal cycler PTC-100TM (MJ Research, Massachusetts). The gel used (4% polyacryla-mide, 5% bisacrylamide) contained 8M Urea. TGGE was performed using TG-180 (Taitec, Japan) for 75 min at 15 V/cm on a temperature gradient of 30–70°C perpendicular to the electric field. The gels were silver stained.

## 2.3. Sequencing

The DNA fragments α, β and γ were extracted from the gel bands, amplified by PCR, cloned into a plasmid using TA cloning kit (Invitrogen) and sequenced using a DSQ 2000L sequencer (Shimadzu, Japan).

## 2.4. Computer analysis

A set of featuring points (~ten) assigned to a genome profile on a computer display, were processed to calculate normalized mobility and temperature of each point, which we call as a species identification dot (spiddo). A measure of similarity of two genomes, pattern similarity score (PaSS) was introduced as follows

$$PaSS = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\left| \bar{P}_i^{(1)} - \bar{P}_i^{(2)} \right|}{\left| \bar{P}_i^{(1)} \right| + \left| \bar{P}_i^{(2)} \right|} \qquad (1)$$

$\bar{P}$ of each spiddo is its position vector and a function of temperature and mobility (i.e. $\bar{P}_i = P(T, m)$). The parenthe-sized superscripts 1 and 2 represent genomes 1 and 2, respectively. PaSS will be unity for a complete match in two sets of spiddos. In general, $0 \leq PaSS \leq 1$.

## 3. Results and discussion

### 3.1. Introduction of spiddos and PaSS

The DNA bands generated by Genome profiling (GP) are representative of the original genome and featuring points expressed on such bands can be used to identify genomes or species (Fig. 1). A set of these points, named species iden-tification dots (spiddos), collected from several DNA bands in a single GP plate can present convenient and useful data. To define spiddos, we established three different protocols: (i) The DNA bands on which spiddos are assigned are defined in advance. (ii) A set of small blocks (patches) on which featuring points are assigned to be spiddos, are
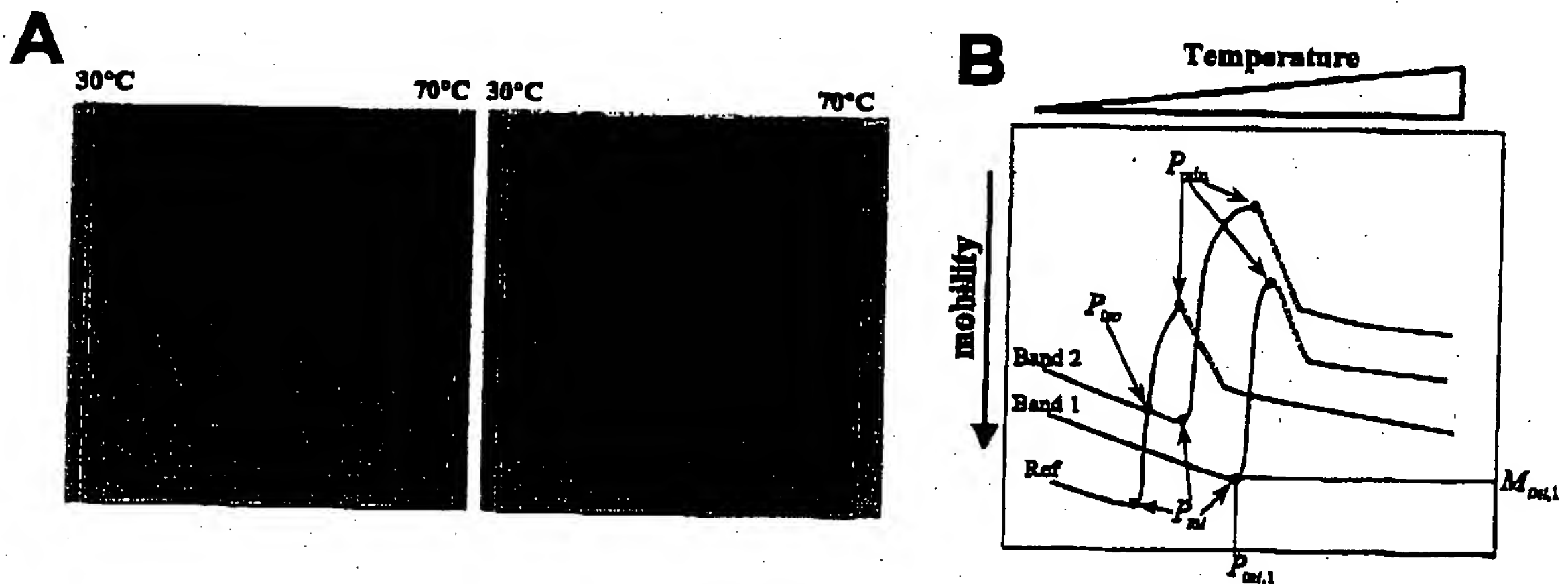


Fig. 1. Reproducibility test of species identification dots (spiddos). Genomic DNA of Saccharomyces cerevisae was examined by Genome profiling (GP), random PCR using a primer (for example, dAGAACGCGCCTG) and subsequent TGGE analysis. (A) Two independent reproducibility tests are shown with spiddos marked. The spiddos of a pair of GPs were processed to calculate the PaSS value (0.991 for the example shown here). Beginning from the DNA extraction, the reproducibility of spiddos was tested ten times (mean, 0.986; SD, 0.003). (B) Schematic presentation of part of the Genome profiling with the spiddos defined. Spiddos are composed of a set of points (about ten) that are extracted from a set of frequently appearing bands representing commonly conserved genetic fragments. Each dot specifies either the initial DNA melting point ($P_{ini}$), the most retarded ($P_{min}$) or an isomobility point of two DNAs ($P_{iso}$). The coordinate of each spiddo is normalized with use of the coordinates of the two points obtained from internal references. The detailed protocol is to be on web (http://gp.fms.saitama-u.ac.jp).

defined in advance. (iii) All possible featuring points (over ten) are extracted as preliminary *spiddos*. A fixed number of dots (such as ten) are taken as actual *spiddos* out of the preliminary ones to calculate the *PaSS* value. Using the same number of dots, the third protocol can be used to compare more distant species. Here we adopt the first protocol and do not further discuss the remaining protocols for clarity. Each *spiddo* is specified by both mobility and temperature, both of which are determined after calibration and normalization of band patterns by a computer using co-migrating internal references (Fig. 1B). Although some reduction of information is inevitable, the resultant *spiddo* profile can provide useful information about a genome and can be used for visual comparison as they are (see Figs. 1 and 2A). Using *spiddos*, the similarity of two genomes can be measured with *pattern similarity score* (*PaSS*) (see Section 2.4). The reliability and reproducibility of *PaSS* has been examined by repeated experiments using yeast genome DNAs, resulting in an average error of 1.4% and a deviation of 0.3% (i.e. *PaSS* = 0.986 ± 0.003) (Fig. 1A). This result is not universal but rather depends on variables such as personnel, apparatus and samples. In other words, experimental accuracy can be increased by using a more rigorous, automated system.

### 3.2. PaSS values obtained from E. coli

Clinical strains of *E. coli* O157: H7 were examined by GP and the subsequent process of obtaining *spiddos* and *PaSS* (Fig. 2A and Table 1). Evidently, all of the *spiddos* prove to be very similar, enabling us to identify them as the same species. On the other hand, sufficiently variant *spiddo* profiles allowed genomic assignment as different strains of *E. coli* (*PaSS* 0.965–0.992; Note that the *PaSS* values were experimentally 0.986 ± 0.003 between the same strains (Fig. 2B)). Since these strains were collected from a relatively small area (Saitama Prefecture, 3800 km²) during the period 1996–7, the closeness of the *PaSS* values between any combination of two strains was reasonable. The *PaSS* values so far obtained between the strains of the same

Table 1
Correlations of the results obtained by three different methods. The italic values (self-correlations) are results obtained for intrafamily members while the bold values are interfamily ones

|        | PaSS            | RFLP            | Sequencing         |
|--------|-----------------|-----------------|--------------------|
| PaSS   | *0.979*<br>**0.979** | 0.253[a]       | − 0.526[a,b]       |
| RFLP   |                 | *0.985*<br>**0.745** | − 0.474[a]         |
| Sequencing |             |                 | *0.0061*[c]<br>**0.0057**[c] |

[a] The degree of correlation can be evaluated from the absolute value.
[b] Only the dots on the bands which were processed to sequencing are used here for comparisons. If the others are included, less correlated (0.208).
[c] Substitutions/nucleotide.

species (*E. coli*, *Saccharomyces cerevisiae*, and *Bacillus subtilis* (data not shown)) were all above 0.95, although this value is not definite but can be used tentatively as a reference for species-identification. *We also obtained a mean PaSS value of significantly greater than 0.70, which is the value for an arbitrary pair of species, for a group of Enterobacteriaceae such as E. coli and Shigella (our unpublished data).*

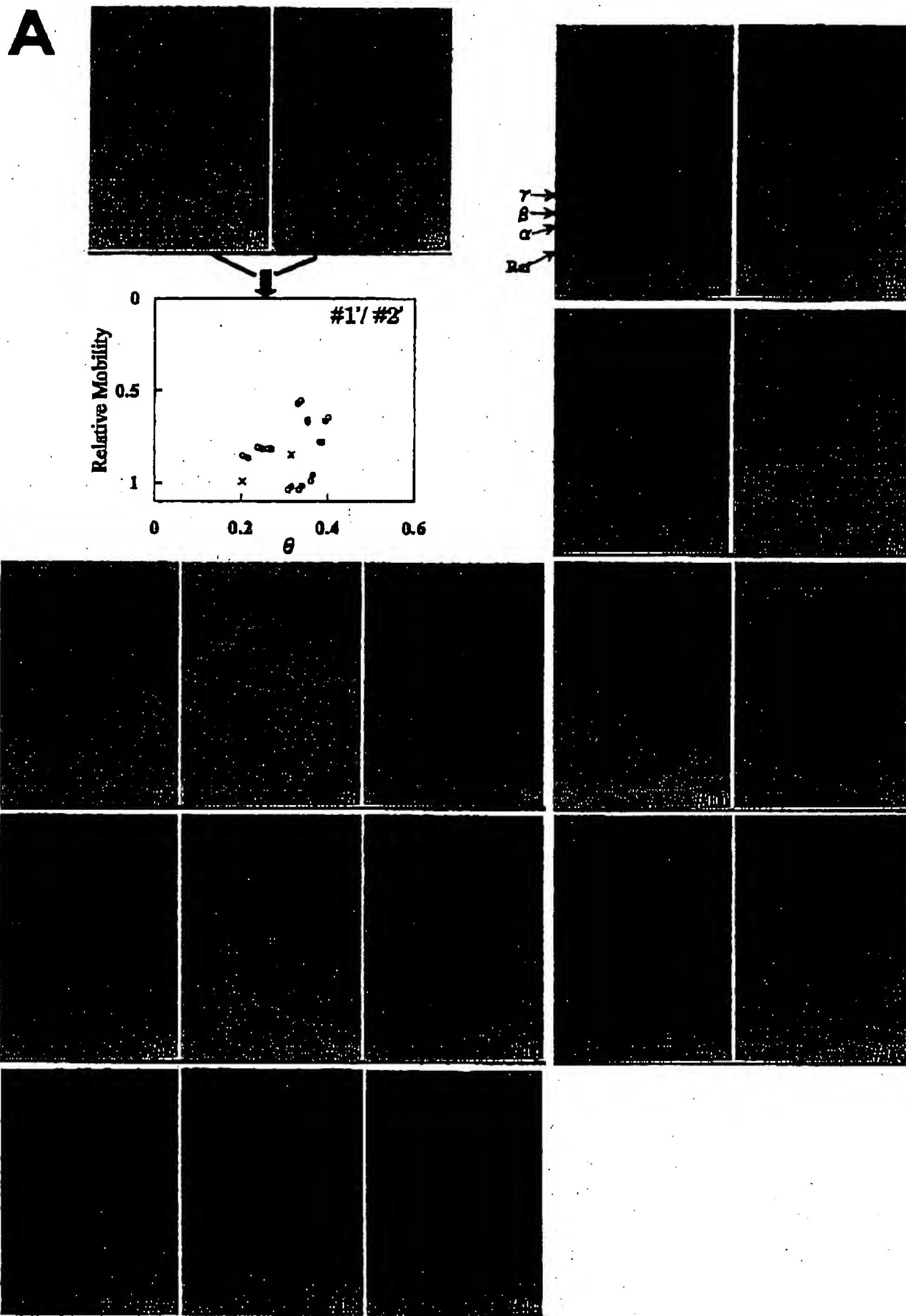### 3.3. PaSS as a measure of genome distance

Here, the nature of the *PaSS* value must be deeply considered. The normalization factors used for both coordinates (mobility and temperature) are essentially independent of each other, offering arbitrariness in weighing the ratio of these two coordinates. These problems of arbitrariness can be solved by empirical approaches in which the parameters are adjusted to give the most rational result. The definition of *PaSS* means that differences in the positions of featuring points expressed in two coordinates can be quantified. The fractional difference ($\Delta X/X$) adopted here is very reasonable for the coordinate of mobility considering that the larger the mobility becomes, the more the difference, but this is rather a convenience for temperature. Although a temperature difference of a particular featuring point between two strains must have been derived from a sequence difference (Nishigaki et al., 2000b), the correlation between the difference in sequence and the shift in the TGGE profile is moderate, but not rigid (Wartell et al., 1990; Steger, 1994). Some point mutations cause a significant shift in the DNA melting temperature whereas others do not (Riesner et al., 1992). Therefore, the relationship between the number of point mutations and the degree of temperature shift is not linear. Nonetheless, statistics points out that more data will improve the relationship between them. Since closeness in the genome sequence and the *PaSS* value are conclusively correlated as discussed above, the *PaSS* value can be used to semi-quantify genomic similarity. Therefore we can draw a hypothetical diagram *(genome space)* for representing relations of species and strains based on *spiddos* and *PaSS* as shown in Fig. 3. We can convert the *PaSS* into a measure of distance, *d*, as follows

$$d = (1 - PaSS/PaSS) \qquad (2)$$

However, the distance, *d*, introduced here does not have the nature of the conventional distance, like $|\vec{d}_{1,3}| = |\vec{d}_{1,2} + \vec{d}_{2,3}|$ (where $\vec{d}_{i,j}$ means position vector), but still have a non-linear additive nature. If we tentatively call it as genome distance, genome distance enables us to imagine '*genome sequence space*' where each genome (in other words, each individual organism) is positioned, each separated from the others by the corresponding 'genome distance'. If *d* is sufficiently small ($d \approx 0$), it means that the two genomes of interest belong to the same species. Genome distance defined here has, in essence, the same meaning with the well-established 'genetic distance' developed by Nei and others (Nei and Takezaki,

1996), though it had been differently defined (genetic distance is defined based on the number of base-substitutions between genes observed at the sequence level). Genetic distance, which is simply defined, has a merit of being unequi-

vocal in its calculation but has a demerit of being practically unhandy (if one would try to obtain sufficiently reliable data, it requires a lot of cloning and sequencing). In this vein, genome distance, which was experimentally introduced,



**A**

#1'/ #2'

Relative Mobility

0

0.5

1

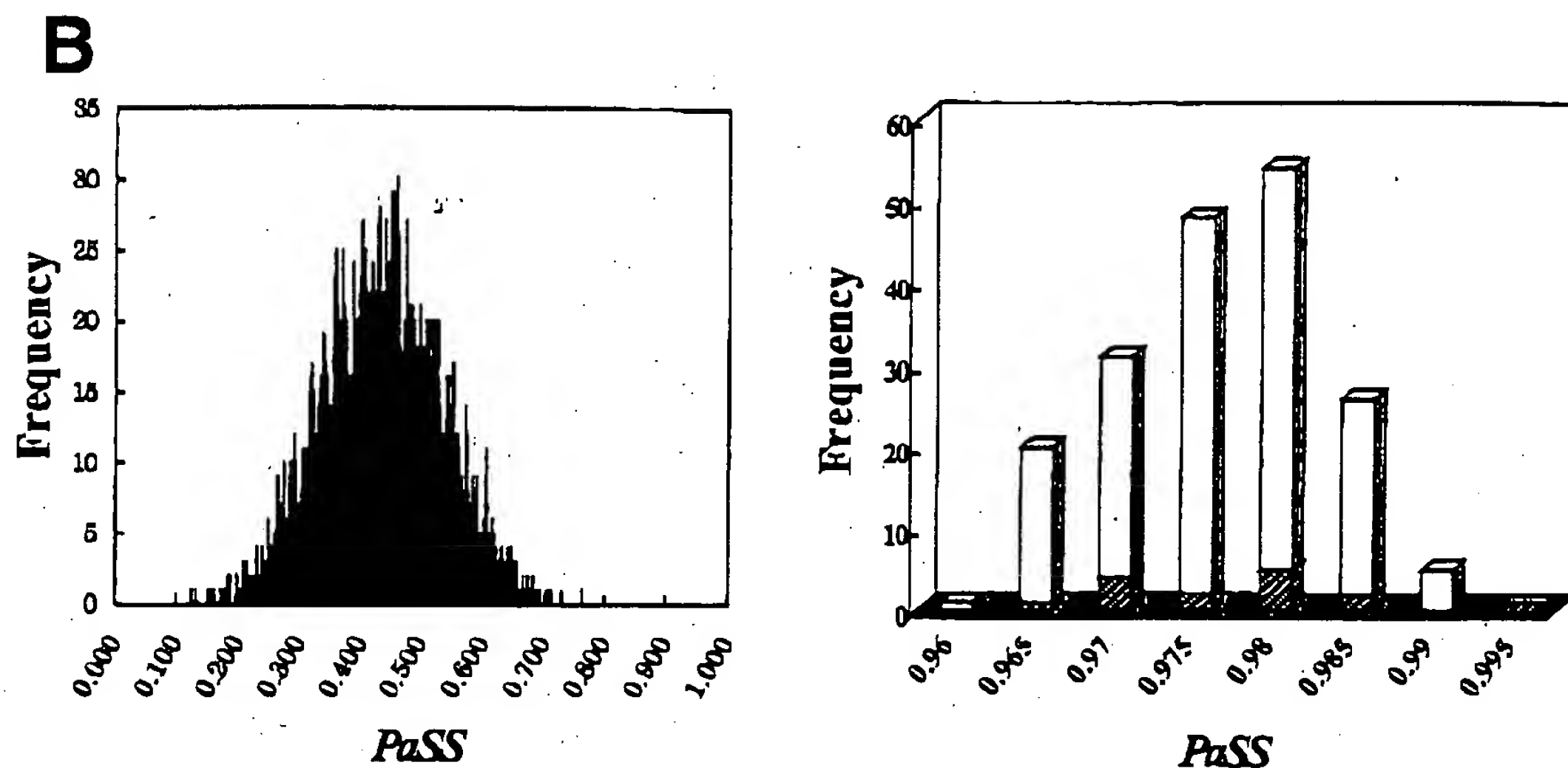0　0.2　0.4　0.6

$\theta$

γ→
β→
α→
Rα→

**B**



Fig. 2. Clinical strains of *E. coli* were analyzed by *Genome profiling*. (A) Genome profiles were obtained using a primer (dAGAACGCGCCTG). Raw GP data and their *spiddos* are shown (#1–19). The computer-normalized *spiddos* are shown on the same plane along the temperature (θ) and mobility axes for #1 and #2 at #1'/#2', where the two points used as the standard points for normalization are shown by the symbol ×. Nineteen genomic DNAs of *E. coli* O157: H7 strains, collected during 1996–7, Saitama prefecture, Japan, are clustered into eight subgroups (tied with a black underline) that were isolated at the same time from members of the same family. Ref refers to the internal reference used and α (265 bp), β (327 bp) and γ (336 bp) are the fragments sequenced. The *PaSS* value for examples #1' and #2' was 0.989. (B) Distribution of pattern similarity scores (*PaSS*). (B; left) *PaSS* distribution for two sets of randomly generated *spiddos*. The average *PaSS* value between them (4950 trials shown) is 0.435 and its SD, 0.091. (B; right) All pairs of the 19 genome profiles were examined and the *PaSS* values for 171 combinatorials were sorted. Those of spatially- and temporally-close pairs (subgroups indicated above) are shown with a hatched box.

can serve as a convenient substitution for genetic distance, although requires a lot of theoretical refinements.

### 3.4. Considerations on methodological differences

We also evaluated our *spiddo*-based technology by comparison with sequencing and RFLP (Restriction Fragment Length Polymorphism) analyses. Fig. 4 shows the sequence and RFLP analysis of the genomic DNAs from

the same set of *E. coli* (19 strains) from which *spiddos* were obtained. Bands α, β, and γ (see Fig. 2A) were sequenced, while the restriction enzyme *Xba 1* products of the whole genome DNA were examined by RFLP. Differences between genomes were identified and quantified by both procedures (partially shown in Fig. 4). Data processing determined a weak-to-moderate correlation between the results obtained by each pair of these methods (Table 1). This is because each method monitors different aspects of
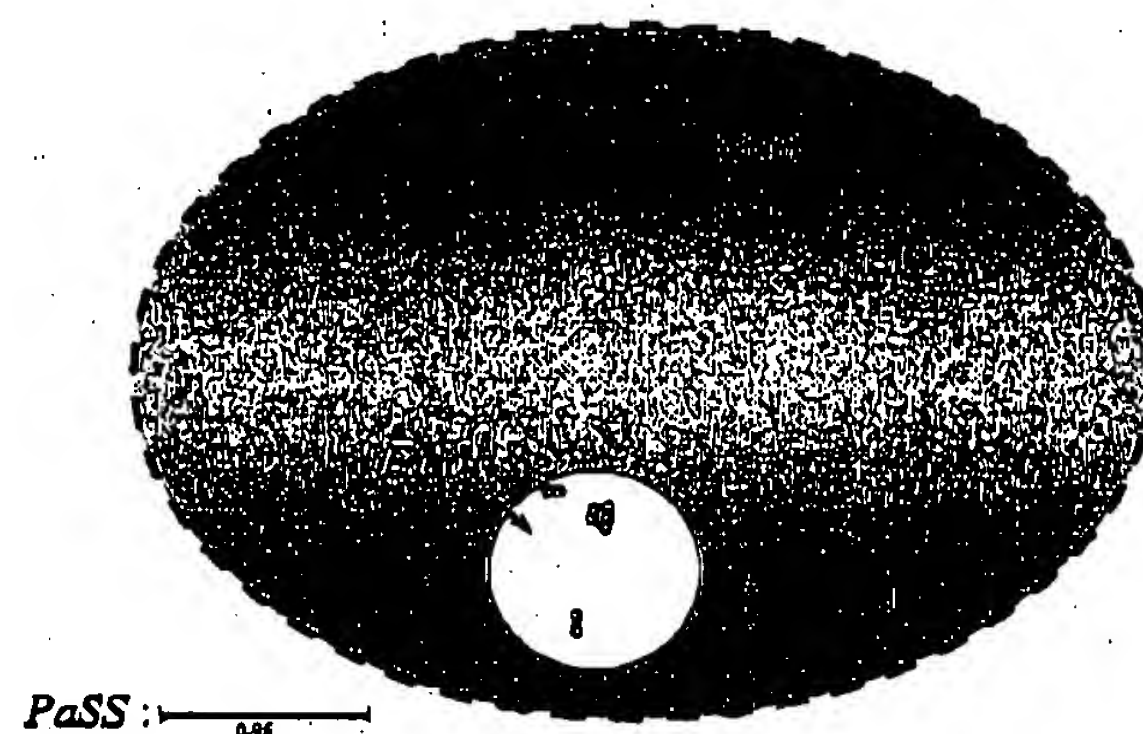


Fig. 3. Hypothetical 'genome space' figured by *spiddos* and *PaSS*. Each genome can be located on this hypothetical *genome space* based on its *spiddos*. For example, the strains examined here are plotted with the smallest circle, which is proportional to experimental error (1.4% of *PaSS*). Within the circle of *E. coli* species each strain is separated from another by the distance measured with the *PaSS* value obtained for the pair of strains. (For convenience, the measure of distance (*d*) can be introduced as follows, although it does not have a canonical nature: $d = (1 - PaSS)/PaSS$.
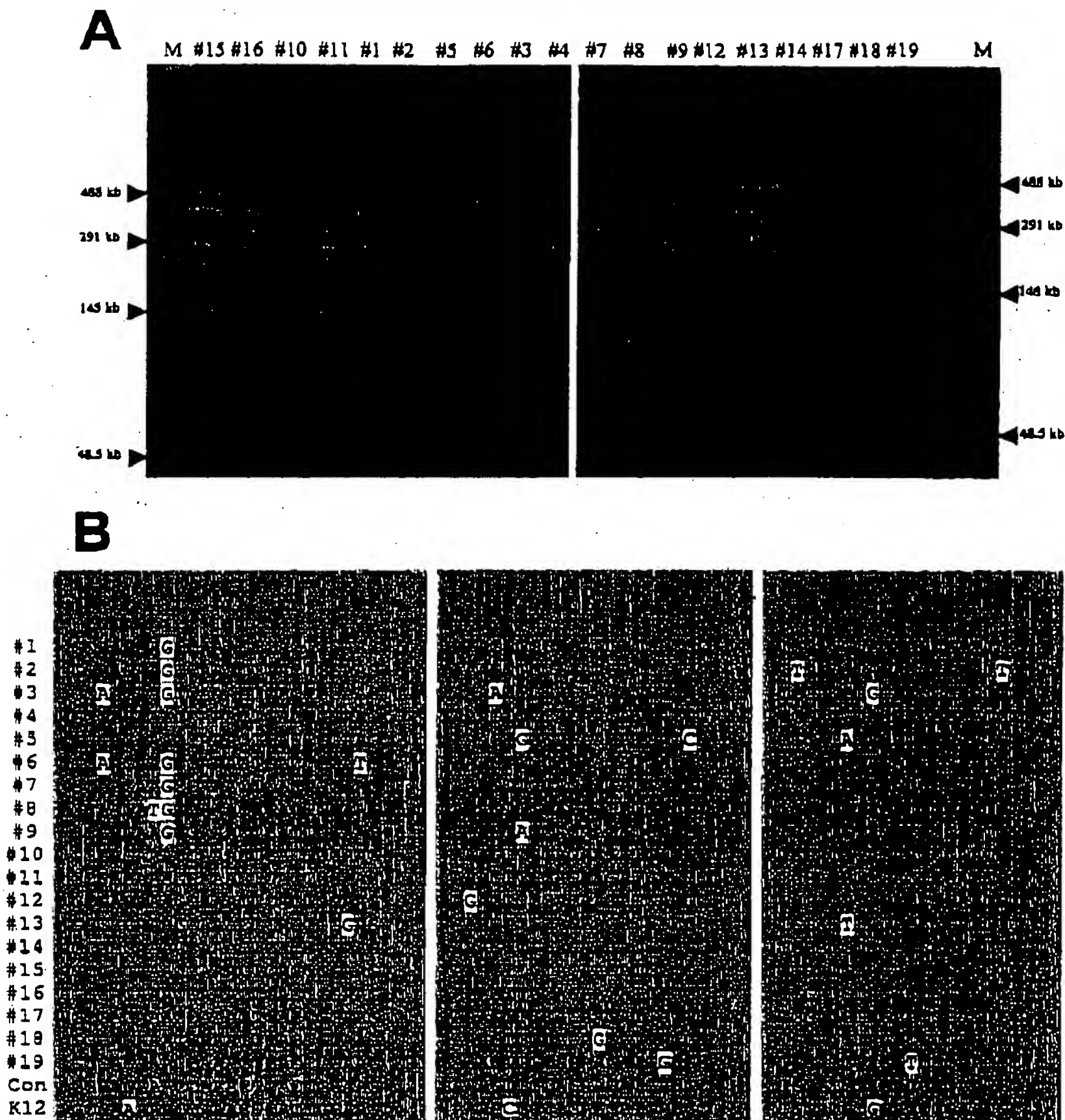
Fig. 4. RFLP and sequencing analyses of clinical strains of *E. coli*. (A) RFLP. Pulsed-field gel electrophoresis analyzed DNA fragments of over 10 kb. The similarity score was calculated as (the number of bands matched/the number of the relevant bands). (B) Sequencing analysis. The DNA fragments, assigned as α, β, and γ in Fig. 2A, were sequenced. Portions of them are presented here. 'Con' and 'K12' mean the consensus sequence among the 19 DNAs examined here and that of *E. coli* K12 that was registered in database, respectively. Throughout this figure, the same numbering for the strains is used as in Fig. 2A.

genomic DNA, of which mutation rates are not identical. This typically appears as a lower correlation (−0.208; normally, −0.526) of sequencing and *PaSS* methods, in case that the *spiddos* are expanded to include those taken from bands that were not sequenced. The disappearance of a particular RFLP band means that nucleotide(s) within the two recognition sites has been altered (see Appendix for further discussion). Therefore, the amount of information obtained is about 1 bit (either mutation-free or not) per band, which is very small. Usually, RFLP by pulsed-field gel electrophoresis, applied to more than 10 kb, is insensi-

tive to deletion/insertion events of 100 bp or less due to the limits of resolution of the electrophoretic technique. This is another downside of RFLP when applied to genome analysis. Sequencing requires pure DNA samples for analysis, a factor that is less important for the other two methods. Sequence data is definitive, but collecting sufficient amounts of it to measure similarity between genomes can be difficult. Moreover, it has already been known that even the very suitable genes for sequence comparison such as *16S rRNA* and *gyrB* often fail to provide PCR products using a same couple of PCR primers *(Yamada et al., 1999)*. And the

amount of information obtained from 16S rRNA is not sufficient to discriminate close species such as *Bacillus cereus*, *B. thuringiensis* and *B. anthracis* (Yamada et al., 1999; Yamamoto and Harayama, 1998). When measuring genome similarity, knowing the entire sequence is not necessary. Rather, differences between genomes should be assigned. In this sense, sequencing may be excessive and not appropriate for the current purpose.

Table 1 also includes a novel result. When the genomes of *E. coli* O157: H7, were classified into two categories [(i), one collected at the same time and same place (spatiotemporal distance ≈ 0) and (ii) the other collected at different times and places (spatiotemporal distance ≫0; months apart or tens of km apart)], all three methods provided rational results. The least sensitive RFLP seems to have generated the most reasonable finding that the similarity of the spatiotemporally-closer genomes was higher (0.985) than that of the more distant genomes (0.745). We have to be sufficiently cautious of this effect, whereby taking a small portion of the whole genome and being too sensitive to it (which seems to be the case with *spiddos* and sequencing methods) may lead to a distorted image. However, only this level of sensitivity can adequately discriminate tiny differences.

The genomic DNA of enterohemorrhagic *E. coli* O157: H7 has been investigated (Kim et al., 1999; Izumiya et al., 1997). One study using pulsed-field gel electrophoresis identified only 6 restriction fragment profiles among all strains collected from various places all over Japan during a period of 1 year (Izumiya et al., 1997). This result can be interpreted to mean that some restriction sites are so well conserved that the overall profiles are essentially the same and that mutations rarely add a difference as shown in Fig. 4A. Obviously, this analysis is effective only for strains with an adequate value, such as 0.1 (that is, 10% of a population lack a particular band), of the product of the mutation rate, $p$, and the number of replications. Suppose a six-cutter is used and $g = 10^4/y$ (for this, more than 24 replications per day throughout a whole year is required, then, the mutation rate, $p$, must be larger than $10^{-5}$, which is much larger than the value commonly believed ($10^{-7 \sim -8}$) for cells cultured under experimental conditions. This means that *notwithstanding such a high mutation rate*, only a very static image of the genome can be observed (rare mutation events confined to some limited regions). Thus, a fine image of genomes cannot be obtained using this approach unless a large amount of electrophoresis is performed.

Other potent methods such as RAPD-PCR (Williams et al., 1990) and Octamer-based genome scanning (Kim et al., 1999) can dig more deeply than RFLP owing to the larger number of manageable fragments generated and the higher sensitivity for point mutations. Both these technologies have the advantage in that they can be easily performed using conventional gel electrophoresis. By changing the primer, these methods can reveal different aspects of genomic sequences. Though not impossible, more elaborate steps

such as Southern blotting are required for RFLP to obtain similar results. Therefore, these methods can be applied to specific purposes such as distinguishing subpopulations of *E. coli* (Kim et al., 1999). However, GP is more advantageous over these methods in being able to use structural stability-based information. Another possible technology for the present purpose must be derived from microarray technology, although it is too costly and too complicated without improvement to be done. Considering all of these, *Genome profiling* reinforced with *spiddos* is, at least currently, the most potent tool for developing genome microbiology.

## 4. Conclusions

1. *Spiddos*, a set of featuring points, could be reproducibly defined from genome profiles. Then *PaSS* was effectively introduced based on *spiddos* to measure the similarity between genomes.
2. *PaSS* could be converted into genome distance ($d$) which easily infers 'genome sequence space' where each genome is separated from the other genomes by $d$. Genome distance, though theoretically juvenile, has a great merit of being easily available.
3. From the methodological comparison, our *spiddo*-based technology proved to provide a purpose-sufficient amount of information; not too much as in the sequencing method or not too less as in RFLP for identification of species.
4. The method developed here allows a shift from phenotype- to genotype-based identification of species.

## Acknowledgements

## Appendix

The vanishing of a band is mainly governed by the mutation event occurred at recognition site. The expectation value, $E_g$, at which we will find a new restriction site (six nucleotides) generated within $n$ nucleotides in length after $g$ rounds of replication, is

$$E_g = \left[ \sum_{m=1}^{6} (1/4)^{6-m}(3/4)^m {}_6C_m(gp)^{6-m}(1 - gp)^m \right] \times n$$

where $p$ is the mutation probability per nucleotide per replication. Let $g$, $p$ and $n$ be equal to $10^4$ generations, $10^{-7}$ and $10^5$ nucleotides, respectively, then $E_g = 5.4 \times 10^{-11}$. This is much smaller than the value, $E_v = 12 \times gp = 1.2 \times 10^{-2}$, at

which we will find the band of interest vanished due to the mutation occurred within the recognition sequences.

## References

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

Hamano, K., Takasawa, T., Kurazono, T., Okuyama, Y., Nishigaki, K., 1996. Genome profiling- establishment and practical evaluation of its methodology. Nikkashi 1996, 54–61.

Izumiya, H., Terajima, J., Wada, A., Inagaki, Y., Itoh, K.-I., Tamura, K., Watanabe, H., 1997. Molecular typing of enterohemorrhagic Escherichia coli O157:H7 isolates in Japan by using pulsed-field gel electrophoresis. J. Clin. Microbiol. 35, 1675–1680.

Kim, J., Nietfeldt, J., Benson, A.K., 1999. Octamer-based genome scanning distinguishes a unique subpopulation of Escherichia coli O157:H7 strains in cattle. Proc. Natl. Acad. Sci. USA 96, 13288–13293.

Nei, M., Takezaki, N., 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144, 389–399.

Nishigaki, K., Amano, K.N., Takasawa, T., 1991. DNA profiling: an approach of systematic characterization, classification, and comparison of genomic DNAs. Chem. Lett. 1991, 1097–1100.

Nishigaki, K., Naimuddin, M., Hamano, K., 2000a. Genome profiling: a realistic solution for genotype-based identification of species. J. Biochem. 128, 107–112.

Nishigaki, K., Saito, A., Hasegawa, T., Naimuddin, M., 2000b. Whole genome sequence-enabled prediction of sequences performed for random PCR products of Escherichia coli. Nucleic Acids Res. 28, 1879–1884.

Riesner, D., Steger, G., Wiese, U., Wulfert, M., Helbey, M., Henco, K., 1992. Temperature-gradient gel electrophoresis for the detection of polymorphic DNA and for quantitative polymerase chain reaction. Electrophoresis 13, 632–636.

Wang, H., Qin, M., Cutler, A.J., 1993. A simple method of preparing plant samples for PCR. Nucleic Acids Res. 21, 4153–4154.

Wartell, R.M., Hosseini, S.H., Morgan, C.P., 1990. Detecting base pair substitutions in DNA fragments by temperature-gradient gel electrophoresis. Nucleic Acids Res. 18, 2699–2705.

Steger, G., 1994. Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction. Nucleic Acids Res. 22, 2760–2768.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., Tingey, S.V., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18, 6531–6535.

Yamada, S., Ohashi, E., Agata, N., Venkateswaran, K., 1999. Cloning and nucleotide sequence analysis of gyrB of Bacillus cereus B. thuringiensis, B. mycoides, and B. anthracis and their application to the detection of B. cereus in rice. Appl. Environ. Microbiol. 65, 1483–1490.

Yamamoto, S., Harayama, S., 1998. Phylogenetic relationships of Pseudomonas putida strains deduced from the nucleotide sequences of gyrB rpoD and 16S rRNA genes. Int. J. Syst. Bacteriol. 48, 813–819.

**FIG. 2**

HUMAN:

CHIMPANZEE:

```
1 CAGACATCTGTGTCCCCCTCAAAAGTCATCCTGCCCCCGGGGGAGGCTCCCTGCTGGTGACA
  CAGACATCTGTGTCCCCCTCAAAAGTCATCCTGCCCCCGGGGGAGGCTCCCTGCTAGGTGACA
   Q  T  S  V  S  P  P  K  V  I  L  P  R  G  G  S  Q  V  T

  TGCAGCACCTCCTGTGACCAGCCCATGTTGTTGGGCCATAGAGACCCCGTTGCCTAAAAAG
  TGCAGCACCTCCTGTGACCAGCCCACGCTTGTTGGGCCATAGAGACCCCGTTGCCTAAAAAG
   C  S  T  S  C  D  Q  P  D  L  L  G  I  E  T  P  L  P  K  K
```

HUMAN:

CHIMPANZEE:

```
121 GAGTTGCTCCTGCCCGGGAACAACCGGAAGGTGTATGAACTGAGCAATGTGCAAGAAGAT
    GAGTTGCTCCTGCCTGGGGGAACAACCGGAAGGTGTATGAACTGAGCAATGTGCAAGAAGAT
     E  L  L  L  G  N  N  W  K  V  Y  E  L  S  N  V  Q  E  D

    AGCCAACCAATGTGCTATTCAAACTGCCCTGATGGGCCAGTCAACAGCTAAAACCTTCCTC
    AGCCAACCAATGTGCTATTCAAACTGCCCTGATGGGCCAGTCAACAGCTAAAACCTTCCTC
     S  Q  P  M  C  Y  S  N  C  P  D  G  Q  S  T  A  K  T  F  L
```

HUMAN:

CHIMPANZEE:

```
241 ACCGTGTACTGGACTCCAGAACGGGTGGAACTGGCCACCCCTCTTGGCCAGCCAGTG
    ACCGTGTACTGGACTCCAGAACGGGTGGAACTGGCCACCCCTCTTGGCCAGCCAGTG
     T  V  Y  W  T  P  E  R  V  E  L  A  P  L  P  S  W  Q  P  V

    GGCAAGAACCTTACCCTTACGCTGCCAGGTGGGCACCCCGGCCAACCTCACC
    GGCAAGAACCTTACCCTTACGCTGCCAGGTGGGCACCCCGGCCAACCTCACC
     G  K  D  L  T  L  R  C  Q  V  E  G  G  A  P  R  A  N  L  T
```

HUMAN:

CHIMPANZEE:

```
361 GTGGTGCTGCTCCGTGGGGAGAAGGAGCTGAAACGGGAGCCAGCTGTGGGGAGCCCGCT
    GTGGTGCTGCTCCGTGGGGAGAAGGAGCTGAAACGGGAGCCAGCTGTGGGGAGCCCGCT
     V  L  R  G  E  K  E  L  K  R  E  P  A  V  G  E  P  A

    GAGGTCACGACCACGGTGCTCTGGTGAGAGAGATCACCATGGAGCCAATTTCTCGTGCCGC
    GAGGTCACGACCACGGTGCTCTGGTGAGAGAGATCACCATGGAGCCAATTTCTCGTGCCGC
     E  V  T  T  T  V  L  V  R  D  H  H  G  A  N  F  S  C  R
```

# Definitions:

Prepared by Dr. Nishigaki

**Random PCR :**

This is a kind of PCR (polymerase chain reaction; which can amplify a specific DNA up to million times) but requires only a single primer (the conventional PCR needs a pair of primers to specify the region of a template DNA to be amplified just as a round trip requires two tickets of "from·to" and the reverse.). What is interesting is that the primer sequence of random PCR can be arbitrary (therefore, this method is also called as *arbitrarily primed PCR*) while the conventional PCR not. However, in this invention it is recommended to select one out of a list of the recommended primers for the sake of generality. This is because the genome profiles (GPs) of organisms to be compared should be obtained by use of the same primer. This method is very useful to sample DNA fragments from the template DNA (genome DNA), which can be used for identifying the organism. The reason why random PCR can sample a variety of fragments from the template DNA comes from the hybridization mechanism of the primer to a template DNA (see below). Owing to this nature, random PCR works just as random sampling in Statistics.

**Hybridization of a random PCR primer to a template DNA:**

Arbitrarily designed primers usually bind to a template DNA in a manner of mismatch·containing hybridization, that is, all of the sequence of the primer cannot be completely complementary to a portion of the template DNA like ;

5' GGCTATGACCTG 3' (primer site)

  * * *   * * *  * *

3' CCGACCCTGATA 5' (template site)

where the asterisk shows Watson·Crick base pairing and 5' and 3' indicate the direction of the DNA strand. So, this example contains three mismatches which lead to less stability of the primer·template complex. The more the mismatches (including bulges), the less stable the complex is. The quantitative measure of the stability had been established based on Thermodynamics (Poland, D., Biopolymers 13,1859·1871(1974)). What is important in this method is that each primer can find its multiple binding sites along any template DNA if the annealing conditions are made less stringent (i.e., lower

temperature) for binding. This is a very simple problem of stochastic; it may be difficult to find a sequence of letters "thermodynamics" (corresponding to a primer sequence) in a particular text (which corresponds to the base sequence of a genome DNA) but more probable if the sequence of letters are relaxed to be " ·the·m···" or " ·the····". Therefore, if the stringency of the annealing conditions is more relaxed, then we can get more species of random PCR products copied from various sites of the template DNA. This is why we can use random PCR as a sampling tool from a genome DNA. Surprisingly, the species and amount of DNA fragments to be amplified have been experimentally shown to be reproducible so far as the template, the primer and the experimental conditions are held to be the same.

## TGGE(/DGGE):

Temperature gradient gel electrophoresis (/Denaturant gradient gel electrophoresis) has been developed to analyze the melting profile of double-/single- stranded DNAs or proteins caused by the effect of temperature (or denaturant). TGGE is recommendable for the sake of GP due to its reproducibility and convenience. DNA melting has been monitored by a spectroscopic measure of UV absorbance which changes depending on the structure of DNA (i.e., single stranded or double stranded). In addition to the simplification of the experimental procedures, the introduction of gel electrophoresis had amazingly decreased the necessary amount of DNA for this purpose by three orders of magnitude (from $10^{-5}$g to $10^{-8}$g). Another important nature of TGGE used for GP is that random PCR products (which are desirable to be multiple because of the necessity of being a large amount of information) can be directly applied to this analysis without separating each DNA fragment. The irreplaceable reason of TGGE is that TGGE can provide the melting profile of DNA (so called a *DNA profile* and collectively *genome profile* herein) which conveys the sequence information. Thus, TGGE can provides both information of the size and the sequence (not the base-sequence itself but the entity related to the sequence through a computer algorithm (Poland, Fixman, Freire 1978?)) of a DNA fragment.

## μ TGGE:

A miniaturized system of TGGE which is substantially important in this methodology since it enables us to complete the whole TGGE process within ten minutes or so with a much reduced amount of samples, resultantly raising the performance by hundreds fold (Biyani et al. Electrophoresis 2001). Though it may not be essential, this technology supports the further development of the whole experimental system of GP.

## DNA melting profile/genome profile:

Double-stranded (ds) DNAs proceed to be partially and then completely unfolded as the temperature is elevated (or the concentration of denaturant such as urea and formamide is raised). In gel, the partially unfolded (or denatured) DNA is known to have a less mobility than that of double-stranded state depending on the extent of denaturation. Therefore, a straight line of DNA layered on the top of a slab gel draw a complicated curve depending on its sequence when it is migrated down into the gel where a straightly-rising temperature gradient is set from left (low) to right (high). The pattern thus obtained is called DNA (melting) profile since the mobility transition appearing here corresponds to the cooperative and partial melting of DNA. Note that the temperature (or the concentration of denaturant) at which mobility transition occurs is governed by the sequence of the DNA (this is why we assert that GP utilizes the sequence information of DNA).When the original solution contains multiple DNA fragments of random PCR products, multiple band patterns appear, which are collectively called as genome profile. This genome profile is proved to be specific to each organism, representing the whole image of genome through random sampling.

## Spiddos (species identification dots):

This term is coined to call a group of characteristic points obtained from a genome profile, most of which represent the beginning point of DNA melting (of which temperature is especially called as $T_{ini}$ ; initial melting temperature). Typically, 8 to 10 points (or spiddos), depending on the number of bands appeared on a gel, are collected for a single Genome profile (GP) for identification of species, though the less spiddos are still usable. These spiddos come one by one from each band, thus representing the same number of DNA fragments. What is technically important is that the spiddos should be intrinsic to each genome. In other words, spiddos should not be arbitrary but specific to the genome DNA and reproducible at any time at any place so far as the same genome of an organism is concerned. This precious nature could be acquired by introducing a normalization process exploiting the common internal references. Characteristic points appearing on a gel are not spiddos by themselves but turn to be spiddos through the normalization process which eliminates experimental fluctuations.

## Normalization process:

It was introduced to eliminate the unavoidable experimental fluctuations common to gel electrophoresis. The most important invention is in an algorithm to find two finite

3

points on a gel, for which both the temperature and the mobility are determined, and to normalize each spiddos by use of these finite points (internal reference points). As explained in *internal reference*, this can be fulfilled by using DNA of a known sequence and a known DNA melting profile. The DNA melting profile can be obtained theoretically based on the sequence of DNA but it is better to confirm it by experiment for the sake of accuracy. Once the DNA melting profile is known, the initial melting point (one of the characteristic points which appear as a bending point of a curve) can be used as a finite reference point, which was established through a number of experiments. Each characteristic point can be called as a spiddo after the normalization as shown below:

$$\mu_i = m_i/m_r$$

$$\theta_i = T_i/T_r$$

where $\mu_i$, $m_i$, and $m_r$ are the normalized mobility of i-th point (spiddo), the experimentally measured mobility of i-th point, and the experimentally measured mobility of the reference point while $\theta_i$, $T_i$, and $T_r$ are the normalized temperature of i-th point (spiddo), the experimentally measured temperature of i-th point, and the experimentally measured temperature of the reference point. Each point adopted here is, in principle, the point which corresponds to the initial melting ($P_{ini}$).

To adopt such a point as $P_{ini}$ is a part of the invention (by no means an ordinary skill in the art).

## PaSS (Pattern similarity score):

This score (and formula) was introduced by us to measure the similarity of two patterns. A Genome profile (GP) of an organism of interest is, first, processed to pick up the characteristic points (i.e., spiddos) from the GP as many as possible. Then, each of those spiddos is made correspondent to one of the spiddos of another organism to be compared in such a manner as the PaSS value defined in the equation XX will take the minimum value among all the possible correspondences. Therefore, this is a kind of stochastic approach. Some spiddos provide a genuine correspondence of two ccgf (commonly conserved genetic fragment) such as homologs and paralogs and the others, an apparent correspondence of no relationship. Even if so, we can expect that the PaSS value thus obtained can represent the closeness between two organisms compared, which has been experimentally demonstrated. If two organisms are close, the PaSS value can be close to 1 (generally, $0 \leqq PaSS \leqq 1$). Therefore, PaSS can be a measure of similarity of any pair of organisms. It is known that the value itself is not linear and not additive, requiring a great care for quantitative handling of them.

**Amount of information required for identification of species:**

As this system is a general approach for genome-based identification of species, it needs to discuss what is the necessary amount of information for identification of species, which was formerly done in our paper. One cannot specify a person in a class only by a criterion whether he/she is wearing glasses or not. Whether the person is taller than the height of 170 cm, whether he/she is heavier than 60 kg, or whether he/she is wearing striped shoes and or else is usually necessary to specify the person uniquely. Then what is the necessary amount of information for this purpose? To solve this problem, it must be questioned what is the entire diversity of species. It is very difficult to answer but can be roughly estimated. One estimation for this is 200 bits equivalent the length of 100 nucleotides. This means that if we can find 100 points of nucleotide in a genome sequence which are quite independent from each other, then we can specify the species uniquely from those nucleotides out of all the possible organisms. It requires not the whole genome sequencing but only a part of it. Genome profiling has been demonstrated to be the very technology which can provide the necessary amount of information for this purpose.

**Sequence-based identification of species:**

This is a rather new movement to try to identify species only by genotype without depending on phenotypic traits. The sequence of 16S/18S rRNA has been used for this purpose and a big database for this purpose has already been built up (Maidak and Woese, RDP II). In this sense, GP is a rather newcomer but it has merits of being simple (it does not do sequencing itself) and abundant in the mount of information (it can add necessary information only by use of different primers). If we obtain GPs probed by each of 5 primers or so, then it is usually sufficient to identify not only species but also individuals uniquely. There are many approaches which try to use sequences of some particular genes just as the 16S/18S rRNA approach. However, currently, none of those are so convenient or sufficient in the amount of information required as the GP approach.

**GP (genome profiling):**

Genome profiling (GP) is a methodology which identifies species based on the genome sequence of an organism. Technically, it consists of three main parts: i) random PCR which has a meaning of collecting several sample DNA fragments from the original genome DNA through the probe of arbitrary sequence of an oligonucleotide (in short,

5

equivalent to random sampling in statistics), ii) TGGE (temperature gradient gel electrophoresis) which serves as an analytical process for the DNA fragments obtained by random PCR and produces a genome profile consisting of multiple characteristic band patterns, which can be changed depending on the sequence of an oligonucleotide used as if one turns over a different page (profile) of a book (genome). iii) Data processing on web or on personal computer which enables us to identify species bases on a database of collected GPs. In the last step, spiddos extracted from a GP has a great role as they represent the genome from which they derived. Throughout this methodology, to use the common probes is recommended so as to make the comparison among all the species possible. Spiddos are normalized and contain the data almost equivalent to the sequence information inherent to an organism.

Internal reference:

In GP, the internal reference has a profound meaning of being used as standard with which the most important normalization is performed. This is the very unique point of GP since no other methodologies seem to claim such requirement as a prerequisite including protein 2D gel electrophoresis (which is more difficult to normalize due to the non-linearity and less reproducibility of pH gradient). The internal reference reports the temperature and the mobility of a definite point (usually, $P_{ini}$; initial melting point) of its band.